Jointly Improving Language Understanding and Generation with Quality-Weighted Weak Supervision of Automatic Labeling

†Ernie Chang, †Vera Demberg, *Alex Marin
†Dept. of Language Science and Technology, Saarland University
*Microsoft Corporation, Redmond, WA
cychang@coli.uni-saarland.de

Abstract

Neural natural language generation (NLG) and understanding (NLU) models are data-hungry and require massive amounts of annotated data to be competitive. Recent frameworks address this bottleneck with generative models that synthesize weak labels at scale, where a small amount of training labels are expert-curated and the rest of the data is automatically annotated. We follow that approach, by automatically constructing a large-scale *weakly-labeled data* with a fine-tuned GPT-2, and employ a semi-supervised framework to jointly train the NLG and NLU models. The proposed framework adapts the parameter updates to the models according to the estimated label-quality. On both the E2E and Weather benchmarks, we show that this weakly supervised training paradigm is an effective approach under low resource scenarios with as little as 10 data instances, and outperforming benchmark systems on both datasets when 100% of training data is used.

1 Introduction

Natural language generation (NLG) is the task that transforms meaning representations (MR) into natural language descriptions (Reiter and Dale, 2000; Barzilay and Lapata, 2005); while natural language understanding (NLU) is the opposite process where text is converted into MR (Zhang and Wang, 2016). These two processes can thus constrain each other – recent exploration of the duality of neural natural language generation (NLG) and understanding (NLU) has led to successful semi-supervised learning techniques where both labeled and unlabeled data can be used for training (Su et al., 2020; Tseng et al., 2020; Schmitt and Schütze, 2019; Qader et al., 2019; Su et al., 2020).

Standard supervised learning for NLG and NLU depends on the access to labeled training data – a major bottleneck in developing new applications.

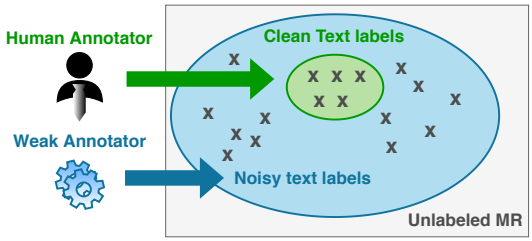

Figure 1: **Training scenario**: Each × represents a labeled data instance. The goal is to learn both from few human-labeled instances (inner) and large amounts of weakly labeled data (outer).

In particular, neural methods require a large annotated dataset for each specific task. The collection process is often prohibitively expensive, especially when specialized domain expertise is required. On the other hand, learning with weak supervision from noisy labels offers a potential solution as it automatically builds imperfect training sets from low cost labeling rules or pretrained models (Zhou, 2018; Ratner et al., 2017; Fries et al., 2020). Further, labeled data and large unlabeled data can be utilized in semi-supervised learning (Lample et al., 2017; Tseng et al., 2020), as a way to *jointly* improve both NLU and NLG models.

To this end, we target a weak supervision scenario (shown in Figure 1) consisting of small, high-quality expert-labeled data and a large set of unlabeled MR instances. We propose to expand the labeled data by automatically annotating the MR samples with noisy text labels. These noisy text labels are generated by a *weak annotator*, which is built upon recent works that directly fine-tune GPT-2 (Radford et al., 2019) on joint meaning representation (MR) and text (Mager et al., 2020; Harkous et al., 2020). Then, we jointly train the NLG and NLU models in a two-step process with semi-supervised learning objectives (Tseng et al., 2020). First, we use pretrained models to estimate quality scores for each sample. Then, we down-weight the

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 818–829
April 19 - 23, 2021.

loss updates in the back-propagation phase using the estimated quality scores. This way, the models are guided to avoid mistakes of the *weak annotator*.

On two benchmarks, E2E (Novikova et al., 2017b) and Weather (Balakrishnan et al., 2019), we utilize varying amount of labeled data and show that the framework is able to successfully learn from the synthetic data generated by *weak annotator*, thereby allowing jointly-trained NLG and NLU models to outperform other baseline systems.

This work makes the following contributions:

1. We propose an automatic method to overcome the lack of text labels by using a fine-tuned language model as a *weak annotator* to construct text labels for the vast amount of MR samples, resulting in a much larger labeled dataset.

2. We propose an effective two-step weak supervision using the dual mutual information (DMI) measure which can be used to modulate parameter updates on the weakly labeled data by providing quality estimates.

3. We show that the approach can even be used to improve upon baselines with 100% data to establish new state-of-the-art performance.

2 Related Work

Learning with Weak Supervision. Learning with weak supervision is a well-studied area that is popularized by the rise of data-driven neural approaches (Ratner et al., 2017; Safranchik et al., 2020; Bach et al., 2017; Wu et al., 2018; Dehghani et al., 2018; Jiang et al., 2018; Chang et al., 2020a; de Souza et al., 2018). Our approach incorporates similar line of work, by providing noisy labels (text) with a fine-tuned LM which incorporates prior knowledge from general-domain text and data-text pair (Budzianowski and Vulić, 2019; Chen et al., 2020; Peng et al., 2020; Mager et al., 2020; Harkous et al., 2020; Shen et al., 2020; Chang et al., 2020b, 2021b,a), and use it as the *weak annotator*, similar by functionality to that of fidelity-weighted learning (Dehghani et al., 2017), or data creation tool *Snorkel* (Ratner et al., 2017).

Learning with Semi-Supervision. Work on semi-supervised learning considers settings with some labeled data and a much larger set of unlabeled data, and then leverages both labeled the unlabeled data as in machine translation (Artetxe et al., 2017; Lample et al., 2017), data-to-text generation (Schmitt and Schütze, 2019; Qader et al., 2019) or more relevantly the joint learning framework for training NLU and NLG (Tseng et al., 2020; Su et al., 2020). Nonetheless, these approaches all assume that a *large collection of text* is available, which is an unrealistic assumption for the task due to the need for expert curation. In our work, we show that both NLU and NLG models can benefit from (1) automatically labeling MR with text, and (2) by semi-supervisedly learning from these samples while accounting for their qualities.

3 Approach

We represent the set of meaning representation (MR) as *X* and the text samples as *Y*. There are no restrictions on the format of the MR: each $x \in X$ can be a set of slot-value pairs, or can take the form of tree-structured semantic definitions as in Balakrishnan et al. (2019). Each text $y \in Y$ consists of a sequence of words.

In our setting, we have (1) k labeled pairs and (2) a large quantity of unlabeled MR set X_U where $|X_U| \gg k > 0$. (We force $k > 0$ as we believe a reasonable generation system needs at least a few demonstrations of the annotation.) This is a realistic setting for novel application domains, as unlabeled MR are usually abundant and can also be easily constructed from predefined schemata. Notably, *we assume no access to outside resources containing in-domain text*. The k annotations are all we know about in-domain text.

The core of our approach consists of first labeling MR samples with text, and then training on the expanded dataset. We start with describing the process of creating weakly labeled data (§4). Next, we delve into the semi-supervised training objectives for the NLU and NLG models, which allow the models to learn from labeled and unlabeled data (§5). Lastly, we explain the training process where NLG and NLU models are jointly optimized in two steps: In *step 1*, we pretrain the models on the weakly-labeled corpus, then *continue updating the models* on the combined data consisting of the weak and real data in *step 2*. Importantly, to account for the noise that comes with the automatic weak annotation, *step 2* trains the model with quality-weighted updates (§6). We depict this process in Figure 2.

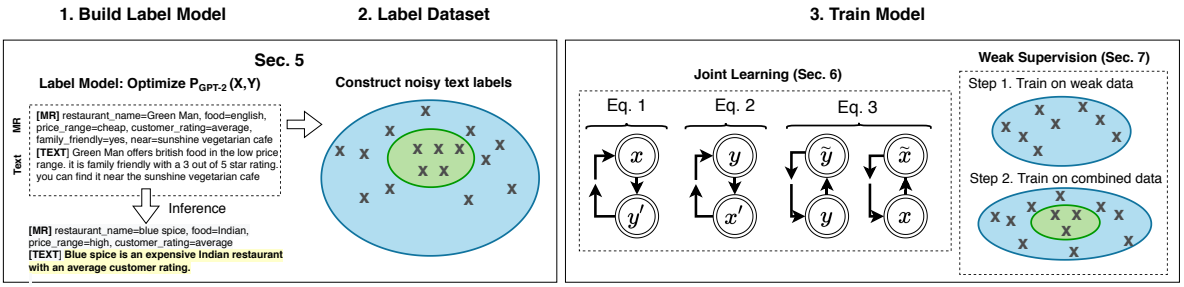

Figure 2: Depiction of the proposed framework. In *joint learning*, gradients are back-propagated through solid lines.

4 Creating Weakly Labeled Data

We construct synthetic data in two ways: (1) creating more MR samples (see §4.1), and (2) by creating a larger parallel set of MRs with texts (see §4.2).

4.1 Generating Synthetic MR Samples

We consider a simple way of MR augmentation via value swapping. This creates more unlabeled MR to be annotated by the *weak annotator* and also provide a substantial augmentation that benefits the autoencoding on MR samples (see Equation 3) by exposing it to a larger set of MR.

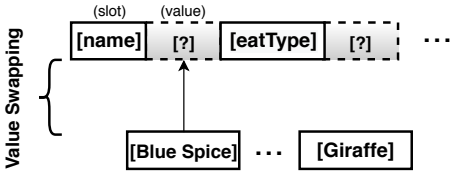

Figure 3: Depiction of MR augmentation in the E2E corpus.

Since each slot in the MR samples corresponds to multiple possible values, we pair each *slot* with a randomly sampled value collected from the set of all MR samples to obtain new combination of slot-value pairs. This way, we create a large synthetic MR set.

4.2 Creation of Parallel MR-to-Text Set

GPT-2 (Radford et al., 2019) is a powerful language model pretrained on the large WebText corpus. Recent work on conditional data-to-text generation (Harkous et al., 2020; Mager et al., 2020) demonstrated that fine-tuning GPT-2 on the joint distribution of MR and text for text-only generation yields impressive performance.

The fine-tuned model generates in-domain text by conditioning on samples from the augmented MR set (X_U). Rather than using GPT-2 outputs directly, we employ them in a process analogous to knowledge distillation (Tan et al., 2018; Tang et al., 2019; Baziotis et al., 2020) where the fine-tuned GPT-2 provides supervisory signals instead of being used directly for generation.

We now describe the process of GPT-2 fine-tuning. Given the sequential MR representation $x_1 \cdots x_M$ and a sentence $y_1 \cdots y_N$ in the labeled dataset (X_L, Y_L), we maximize the joint probability $p_{\text{GPT-2}}(X_L, Y_L)$, where each sequence is concatenated into "[MR] $x_1 \cdots x_M$ [TEXT] $y_1 \cdots y_N$". In addition, we also freeze the input embeddings when fine-tuning had positive impact on performance, following Mager et al. (2020). At test time, we provide the MR samples as context as in conventional conditional text generation:

$$\tilde{y}_j = \arg\max_{y_j} \{p_{\text{GPT-2}}(y_j \mid y_{1:j-1}, x_{1:N})\}$$

The fine-tuned LM conditions on augmented MR sample set X_U to generate the in-domain text[1], forming the weak label dataset $\mathcal{D}_W = (X_U, \tilde{Y}_L)$ with noisy labels $\tilde{y}_i \in \tilde{Y}_L$. In practice, the fine-tuned LM produces malformed, synthetic text which does not fully match with the MR it was conditioned on, as it might hallucinate additional values not consistent with its MR counterpart. Thus, it is necessary to check for factual consistency (Moryossef et al., 2019). We address this point next.

Past findings showed (e.g. (Wang, 2019)) that the removal of utterance with "hallucinated" facts (MR values) from MR leads to considerable performance gain, since inconsistent MR-Text correspondence might misguide systems to generate incorrect facts and deteriorate the NLG outputs. We filter out the synthetic, poor quality MR-text

[1]We adopt the Top-k random sampling with $k = 2$ to encourage diversity and reduce repetition (Radford et al., 2019)

pairs by training a separate NLU model on the original labeled data to predict MR from generated text labels. These MRs can then be checked against the paired MR in $\mathcal{D}_W$ via pattern matching as inspired by Cai and Knight (2013); Wiseman et al. (2017). Specifically, we use a measure of semantic similarity in terms of *f-score via matching of slots between the two MRs*. We keep all MR-text pairs with f-scores above 0.7, as we found empirically that this criterion retains a sufficiently large amount of high-quality data. The removed text sentences are used for unsupervised training objectives as in Eq. 1-3. Using this method, we create a collection of parallel MR-text samples (~500k) an order of magnitude larger than even the full training sets (~40k for E2E and ~25k for Weather).

5 Joint learning of NLG and NLU

For both NLU and NLG models, we adopt the same architecture as Tseng et al. (2020), which use two Bi-LSTM-based (Hochreiter and Schmidhuber, 1997) encoders for each model. The NLU decoder for slot-value structured data (e.g., E2E, Mrkšić et al., 2017) contains several 1-layer feed-forward neural classifiers for each slot; while for tree-structured meaning representation in Balakrishnan et al. (2019), the decoder is LSTM-based. In this framework, both NLU and NLG models are trained to infer the shared latent variable repeatedly – starting from either MR or text, in order to encourage semantic consistency. Each model can be improved via gradient passing between them using REINFORCE (Williams, 1992). This way, the models benefit from each other's training in a process known as the *dual learning* (Su et al., 2020), which consists of both *unsupervised* and *supervised* learning objectives. We now go into details describing them.

Unsupervised Learning. Starting from either a MR sample or a text sample, the models project the sample from one space into the other, then map it back to the original space (either MR or text sample, respectively), and compute the reconstruction loss after the two operations. This repetition will result in aligned pairs between the MR samples and corresponding text (He et al., 2016). Specifically, let $p_\theta(y|x)$ be the probability distribution to map x to its corresponding y (NLG), and $p_\phi(x|y)$ be the probability distribution to map y back to x (NLU). Starting from $x \in X$, its objective is:

$$\max_{\phi} \mathbb{E}_{x\sim p(X)} \log p_\phi(x|y');\ y' \sim p_\theta(y|x) \quad (1)$$

which ensures the semantic consistency by first performing NLG accompanied by NLU in direction $x \rightarrow y' \rightarrow x$. Note that only p_ϕ is updated in this direction and p_θ serves only as as an auxiliary function to provide pseudo samples y' from x. Similarly, starting from $y \in Y$, the objective ensures semantic consistency in the direction where the NLU step is followed by NLG: $y \rightarrow x' \rightarrow y$[2]:

$$\max_{\theta} \mathbb{E}_{y\sim p(Y)} \log p_\theta(y|x');\ x' \sim p_\phi(x|y) \quad (2)$$

We further add two autoencoding objectives on both MR and text samples:

$$\max_{\theta,\phi} \mathbb{E}_{x\sim p(X), y\sim p(Y)} \log p_\phi(x|x) p_\theta(y|y) \quad (3)$$

Thus, unlabeled text samples can be used as they are shown to benefit the text space (Y) by introducing new signals into learning directions $y \rightarrow x' \rightarrow y$ and $\tilde{y} \rightarrow y$. Thus, *we use all in-domain text data* whether they have corresponding MR or not. Note that following (Tseng et al., 2020), we also adopt the variational optimization objective upon the latent variable z which was shown to pull the inferred posteriors $q(z|x)$ and $q(z|y)$ closer to each other. In this case, the parameters of both NLG and NLU models are updated.

Supervised Learning. Apart from the above unsupervised objectives, we can impose the supervised objective on the k labeled pairs:

$$\max_{\theta,\phi} \mathbb{E}_{x,y\sim p(X_L,Y_L)} \log p_\theta(y|x) + \log p_\phi(x|y) \quad (4)$$

Each MR is flattened into a sequence and fed into the NLG encoder, giving NLG and NLU models an inductive bias to project similar MR/text into the surrounding latent space (Chisholm et al., 2017). As we observed anecdotally[3], the information flow enabled by REINFORCE allows the models to utilize unlabeled MR and text, boosting the performance in our scenarios.

[2] This direction is usually termed as *back translation* in MT community (Sennrich et al., 2016; Lample et al., 2018)

[3] Tseng et al. (2020) noticed similar trend in the experiments.

6 Learning with Weak Supervision

The primary challenge that arises from the synthetic data is the *noise* introduced during the generation process. Noisy and poor quality labels tend to bring little to no improvements (Elman, 1993; Frénay and Verleysen, 2013). To better train on the large and noisy corpus described in section §4 (size ~500k), we employ a two-step training process motivated by fidelity-weighted learning (Dehghani et al., 2018). The two-step process consists of (1) *pretraining* and (2) *quality-weighted fine-tuning* to account for the heterogenous data quality.

Step 1: ***Pre-train two sets of models on weak and clean data, respectively.*** We train the first set of models (*teacher*) consisting of NLU, NLG, and autoencoder (AUTO) models on the clean data. The second set of models (i.e. NLU and NLG) is the *student* that pretrains on the weak data.

Step 2: ***Fine-tune the student model parameters on the combined clean and weak datasets.*** We use each *teacher model* to determine the step size for each iteration of the stochastic gradient descent (SGD) by down-weighting the training step of the corresponding *student model* using the sample quality given by the teacher. Data points with true labels will have high quality, and thus will be given a larger step-size when updating the parameters; conversely, we down-weight the training steps of the student for data points where the teacher is not confident. For this specific fine-tuning process, we update the parameters of the student (i.e. NLG and NLU models) at time t by training with SGD, where $\mathcal{L}(\cdot)$ is the loss of predicting $\hat{y}$ for an input x_i when the label is $\tilde{y}$. The weighted step is then $c(x_i, \tilde{y}_i)\nabla\mathcal{L}(\hat{y}, \tilde{y})$, where $c(\cdot)$ is a scoring function learned by the *teacher* taking as input MR x_i and its noisy text label $\tilde{y}_i$. In essence, we control the degree of parameter updates to the *student* based on how reliable its labels are according to the *teacher*.

We denote $c(\cdot)$ as the function of the label quality based on the *dual mutual information (DMI)*, defined as the *absolute* difference between mutual information (MI)[4] in inference directions $x \rightarrow y$ and $y \rightarrow x$. Bugliarello et al. (2020) shows that $\text{MI}_{x\rightarrow y}$ correlates to the difficulty in predicting y from x, and vice versa. Thus we expect the difference between $\text{MI}_{x\rightarrow y}$ and $\text{MI}_{y\rightarrow x}$ for clean sample (x, y) to be relatively small compared to noisy samples, since the level of difficulty is largely *proportional* between NLU and NLG on the samples – difficulty in inferring x from y will result in harder prediction of y from x. Based on this intuition, the DMI score of the sample (x, y) is defined as:

$$exp\left\{\left|\underbrace{\log\frac{q_{\text{AUTO}}(y)}{q_{\text{NLG}}(y|x)}}_{\text{MI}_{x\rightarrow y}} - \underbrace{\log\frac{q_{\text{AUTO}}(x)}{q_{\text{NLU}}(x|y)}}_{\text{MI}_{y\rightarrow x}}\right|\right\}.$$

where $q(\cdot)$ are the two respective models. The DMI for a clean MR-text pair should be *relatively small*, as the two sides contain proportional semantic information[5], and so poor quality samples tend to have higher DMI scores and lower $c(\cdot)$ as they are *less semantically aligned*. Thus, $c(\cdot)$ defines the *confidence* (quality) the teacher has about the current MR-text sample. We use $c(\cdot)$ to scale η_t. Note that $\eta_t(t)$ does not necessarily depend on each data point, whereas $c(\cdot)$ does. We define $c(x_t, y_t)$ as:

$$c(x_t, y_t) = 1 - \mathcal{N}(\text{DMI}(x_t, y_t))$$

where $\mathcal{N}(\cdot)$ normalizes DMI over all samples in both clean and weak data to be in $[0, 1]$.

7 Experiment Setting

Data. We conduct experiments on the Weather (Balakrishnan et al., 2019) and E2E (Novikova et al., 2017b) datasets. Weather contains 25k instances of tree-structure annotations. E2E is a crowd-sourced dataset containing 50k instances in the restaurant domain. The inputs are dialogue acts consisting of three to eight slot-value pairs.

Configurations. Both NLU and NLG models are implemented in PyTorch (Paszke et al., 2019) with 2 Bi-LSTM layers and 200-dimensional token embeddings and Adam optimizer (Kingma and Ba, 2014) with initial learning rate at 0.0002. Batch size is kept at 28 and we employ beam search with size 3 for decoding. The score is averaged over 10 random initialization runs. In our implementation, the sequence-to-sequence models are built upon the bi-directional long short-term memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997). For LSTM cells, both the encoder and decoder have 2 layers, amounting to 18M parameters for

[4] Mutual information for $x \rightarrow y$ can be seen as $H(x \rightarrow y) = H_{AUTO}(y) - H_{NLG}(y|x)$ (Bugliarello et al., 2020).

[5] We found that mutual information for $x \rightarrow y$ is usually greater than that of $y \rightarrow x$ since NLG is a one-to-many and more difficult process as opposed to NLU.

Model	E2E (NLG)					E2E (NLU)				
	10	50	1%	5%	50%	10	50	1%	5%	50%
WA	0.195	0.287	0.563	0.649	0.714	9.48	11.66	13.20	45.21	65.81
JUG*	0.002	0.015	0.726	0.7671	0.819	0.00	0.00	32.24	53.20	**78.93**
decoupled	0.261	0.279	0.648	0.693	0.793	0.00	0.00	20.51	52.77	73.68
joint	0.218	0.336	0.732	0.764	0.775	0.00	6.18	24.98	49.66	70.33
joint+aug	0.275	0.381	0.748	0.781	0.797	5.88	15.79	25.15	53.20	69.68
step 1	0.441	0.487	0.610	0.642	0.685	13.18	14.28	15.37	44.72	65.20
Ours (step 1+2)	**0.489**	**0.558**	**0.754**	**0.775**	**0.822**	**15.81**	**23.67**	**34.09**	**56.33**	72.45

Model	Weather (NLG)					Weather (NLU)				
	10	50	1%	5%	50%	10	50	1%	5%	50%
WA	0.261	0.332	0.518	0.567	0.611	8.42	30.64	66.41	70.19	75.26
JUG*	0.005	0.244	0.618	0.670	**0.726**	0.00	33.48	67.44	79.19	**89.17**
decoupled	0.250	0.288	0.598	0.632	0.719	0.00	28.21	70.24	73.46	88.45
joint	0.270	0.348	0.577	0.639	0.658	0.00	24.52	64.30	69.92	86.86
joint+aug	0.329	0.361	0.589	0.662	0.671	4.21	26.33	67.43	71.19	87.10
step 1	0.371	0.429	0.570	0.607	0.632	12.19	35.89	72.90	72.01	84.73
Ours (step 1+2)	**0.401**	**0.458**	**0.644**	**0.672**	0.717	**16.62**	**42.74**	**75.94**	**80.36**	87.77

Table 1: Performance for NLG (BLEU-4) and NLU (joint accuracy (%)) on E2E and Weather datasets with increasing amount of labeled data from 10, 50 labeled instances to 1%, 5%, and 100% of the labeled data (D_L). Models that have access to **unlabeled ground-truth text labels** are marked with *. We provide results for the NLG and NLU models trained separately using supervised objectives alone (*decoupled*), our semi-supervised joint-learning model (*joint*), *joint* with all unlabeled data (*joint+aug*), and weakly-supervised models (*step 1*). *Step 1+2* denotes the full proposed approach.

	D_L	D_W	X_U	Y_{SL}	Y_{WL}
JUG	✓	✗	✓	✓	✗
WA	✓	✗	✗	✗	✗
decoupled	✓	✗	✗	✗	✗
joint	✓	✗	✓	✗	✗
joint+aug	✓	✗	✓	✗	✓
step 1	✗	✓	✓	✗	✓
Ours (step 1+2)	✓	✓	✓	✗	✓

Table 2: Summary of training data used in each model. Sources of data include labeled data (D_L), unlabeled MR (X_U), weakly labeled data (D_W), 100% real text (Y_{SL}), and weak text labels (Y_{WL}).

E2E NLG	**BLEU-4**
TGEN (Dušek and Jurcicek, 2016)	0.6593
SLUG (Juraska et al., 2018)	0.6619
Dual supervised learning (Su et al., 2019)	0.5716
JUG (Tseng et al., 2020)	0.6855
GPT2-FT (Chen et al., 2020)	0.6562
WA (Harkous et al., 2020)	0.6445
Ours (step 1+2)	**0.7025**

Weather NLG	**BLEU-4**
S2S-CONSTR (Balakrishnan et al., 2019)	0.7660
JUG (Tseng et al., 2020)	0.7768
Ours (step 1+2)	**0.7986**

Table 3: For comparison, we show the performance of previous systems on the datasets following the **original split**, so the scores are **not comparable** to Table 1.

the seq2seq model. All models were trained on 1 Nvidia V100 GPU (32GB and CUDA Version 10.2) for 10k steps. The average training time for seq2seq model was approximately 1 hour, and roughly 2 hours for the proposed semi-supervised training with 100% data. The total number of updates is set to 10k steps for all training and patience is set as 100 updates. At decoding time, sentences are generated using greedy decoding.

8 Results

We first compare our model with other baselines on both datasets, then perform a set of ablation studies on the E2E dataset to see the effects of each component. Finally, we analyze the strength of the *weak annotator*, and the effect of the quality-weighted weak supervision, before concluding with the analysis of dual mutual information.

In particular, we experiment with various low resource conditions of training set (10 instances, 50 instances, 1% of all data, 5% of all data). To show that our proposed approach is consistently better, we include the scenario with 0-100% of the data at 10% interval, to show that performance does not deteriorate as more training samples are added (Figure 4). Table 2 shows the summary of training data used for all models in Table 1. We compare our model with (1) a fine-tuned GPT2 model (GPT2-FT) that uses a switch mechanism to select between input and GPT2 knowledge (Chen et al., 2020)[6], (2) a fine-tuning approach to be used as the weak

[6] https://github.com/czyssrs/Few-Shot-NLG

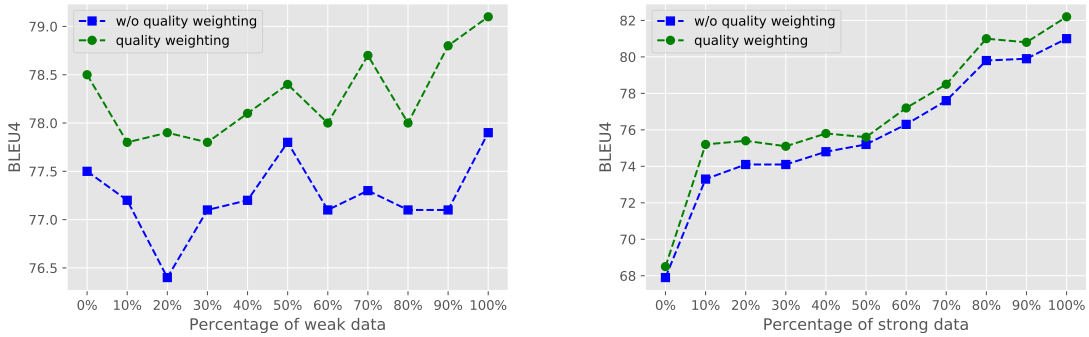

Figure 4: Model performance (BLEU-4) on 5% E2E data with varying percentages of *strong* and *weak* data *with* and *without* DMI-based quality weighting. **Left plot** begins with models trained on labeled data while **right plot** starts with the weak synthesized dataset instead.

Method	NLU			NLG		
	Miss	Redundant	Wrong	Fluency	Miss	Wrong
decoupled	72	78	87	4.10	69	73
JUG	65	**72**	75	4.23	64	65
Ours (step 1+2)	**54**	77	**68**	**4.50**	**63**	**61**

Table 4: Human evaluation on the sampled E2E outputs (100 instances) for models with 1% training data. Numbers of *missing*, *redundant* and *wrong* predictions on slot-value pairs are reported for NLU; *fluency*, numbers of *missing* or *wrong* generated slot values are listed for NLG.

annotator (WA) that predict text from MR or MR from text, depending on the input format during fine-tuning (Harkous et al., 2020)[7], and (3) the semi-supervised model [8] (JUG) from Tseng et al. (2020). Note that the specialized encoder in *GPT2-FT* cannot be easily adapted to the tree-structured input in Weather, and so we do not provide its score on the Weather dataset.

In Table 1, we show that our proposed approach (*step 1+2*) generally performs better than the baselines for both tasks (NLG and NLU) for most selected labeled data sizes. We show that even with only 10 labeled instances, our approach (*step 1+2*) is able to yield decent results compared to the baselines. The difference between models tends to be larger for settings with few training instances, and the advantage of the method diminishes *as the amount of labeled data available for* JUG *increases*, to the point where JUG is able to outperform the proposed approach. Overall, the benefit of the noisy supervisory signal from the weak data is able to boost performance, especially at lower resource conditions.

We observe that training with weakly labeled data alone (step 1) is not sufficient, and so strong data is required to provide the supervisory signals necessary (step 2). Further, the fact that *joint+aug* displays noticeable improvements over *joint* suggests that simply having augmented text helps to improve the encoded latent space as projected by both the NLU and NLG encoders. This also shows an alternative way to introduce additional in-domain information to both models, even though the NLU model does not benefit directly from additional text. Importantly, our approach shows that the *weak annotator* is able to bridge the gap as defined by the access to ground-truth text labels in JUG – outperforming it significantly at low resource conditions (10, 50, 1%, 5%) with the difference in NLG being as large as 48.7 BLEU points with 10 instances. We find that the proposed model also performs well in the high resource (100% of labeled data) condition, as shown in Table 3. Moreover, with 100% labeled data, our model is still able to produce superior performance over some of the baselines, which shows that weak annotation does capture additional useful patterns that benefit the NLG process.

9 Analysis

Error Analysis. Since word-level overlapping scores usually correlate rather poorly with human judgements on fluency and information accuracy (Reiter and Belz, 2009; Novikova et al., 2017a), we perform human evaluation on the E2E corpus on 100 sampled generation outputs. For each MR-text pair, the annotator is instructed to evaluate the *fluency* (score 1-5, with 5 being most fluent), *miss* (count of MR slots that were missed) and *wrong* (count of included slots not in MR) are presented in Table 4, where fluency scores are averaged over 50 crowdworkers. We show that with 1% data, both NLU and NLG models yield signif-

[7] No released source code so we re-implemented it based on paper.

[8] https://github.com/andy194673/Joint-NLU-NLG

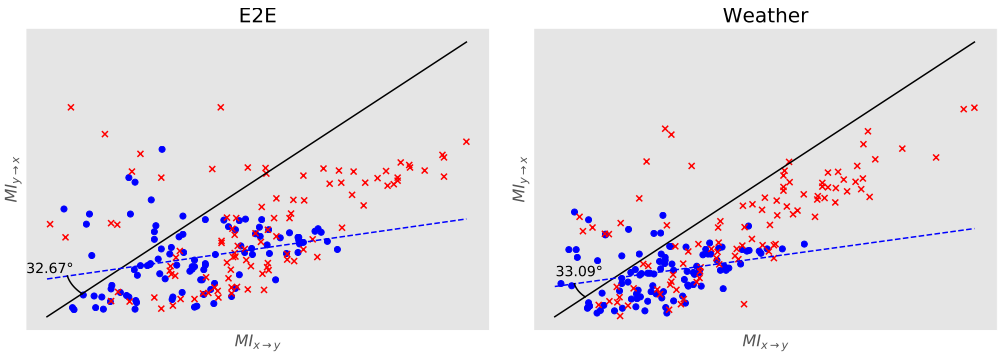

Figure 5: Visualization of dual mutual information (DMI) on both datasets where × markers are 50 random samples from annotated data and ○ markers are 50 random samples in the weak dataset. **Dotted** lines are trend lines for ○ markers and **solid black** lines are diagonal reference that correspond to the perfect NLG-NLU balance where both tasks have equal difficulties.

Method	NLG BLEU-4 (Accuracy (%))	NLU Accuracy (%) (F1)
w/ *DF*	0.683 (77.69)	24.71 (0.6443)
with *DF*	0.703 (79.08)	27.19 (0.6840)
with *WS*	0.733 (82.65)	30.23 (0.7028)
with *WS+CW*	**0.754 (86.44)**	**34.09 (0.7200)**

Table 5: Ablation study of weak supervision (1% E2E labeled data D_L) including data fidelity (*DF*), the proposed model (step 1+2) with weak supervision (*WS*), and *WS* with quality-weighted weak supervision (*WS+CW*).

icantly fewer errors in terms of *misses* and *wrong* facts, while having more fluent outputs. However, it generates more redundant slot-value pairs which we attribute to the noisy augmentation that "misguided" the NLU model.

How Strong is the *Weak Annotator*? To assess the strength of the weak annotator (WA) itself, we also computed its NLG scores with varying amounts of labeled data (see Table 1). We observe that the WA suffers from a performance drop in lower resource conditions (i.e. 0.195 BLEU with 10 labeled instances), when the given training samples are not sufficient for the pretrained model to converge upon a region of in-domain generation. However, it yields some quality data when conditioned on a large number of possible MR (i.e. 50% data), forming a useful in-domain text set (See Table 6).

Analysis of Weak Supervision. In Table 5, we present the results of an ablation study on weak supervision (see §6) where the effect of *data fidelity* is stronger on NLU than on NLG, which is due to the nature of the filtering process which removes faulty text labels which influences both $x \to y$ and $y \to y$ training directions. Next, though weak supervision boosted the model by giving direct supervision in training directions $x \to y$ and $y \to x$, the noisy nature of the augmentation limits its effectiveness. The model is further improved with the proposed quality-weighted update that takes into account the sample quality and alleviate the influence of poor quality samples. Refer to Table 7 for output comparison.

Analysis of the Two-Step Training Process. As inspired by Dehghani et al. (2018), we justify the two-step training process by performing two types of experiments with 5% data (see Figure 4): In the first experiment, we use all the available strong data but consider different ratios of the entire weak dataset – as used in our 2-step approach. In the second, we fix the amount of weak data and provide the model with varying amounts of strong data. The results show that the *student* models are generally better off by having the *teacher*'s supervision. Further, pretraining on weak data prior to fine-tuning on strong data appears to be the better approach and this motivates the reasoning behind our two-step approach.

Analysis of the Dual Mutual Information. Figure 5 depicts DMI with the visualization of $MI_{x\to y}$ as x-axis and $MI_{y\to x}$ as y-axis, in which 100 randomly sampled noisy and ground-truth samples are plotted for both datasets. On the plot, the diagonal reference represents the scenario in which NLG and NLU inference are equally difficult, and we see that annotated data cluster more around the diagonal reference. This means that expert-labeled samples' DMI scores tend to be smaller, where NLU and NLG inference for these samples carry similar levels of difficulty. Importantly, since DMI scores are normalized over both clean and noisy samples, *the proximity of data to the trendlines can then be used to estimate the sample quality* – clean data are closer as compared to the noisy sam-

mr synthetic reference	[name] Giraffe, [eat type] pub, [area] riverside Giraffe is a pub in the riverside of the city just down the street.
mr synthetic reference	[name] Strada, [eat type] restaurant, [food] Italian, [area] city centre, [familyfriendly] no, [near] Avalon Strada is an Italian restaurant not for the families! it is near Avalon in the city centre.
mr synthetic reference	[name] Cocum, [eat type] restaurant, [food] French, [area] riverside, [familyfriendly] no, [near] Raja Indian Cuisine Cocum sells French food near Raja Indian Cuisine.

Table 6: Display of weakly-labeled data samples.

mr	[name] Blue Spice, [eat type] coffee shop, [area] city centre
step 1+2	Blue Spice is a coffee shop in the city centre that of the city.
JUG	Blue Spice serves Italian food and is family friendly.
decoupled	Blue Spice is an adult Italian coffee shop with high customer rating located in

Table 7: Display of text generations from different models.

ples. Thus clean data will have smaller normalized scores, higher $c(\cdot)$, and a larger update step. This further supports the use of the proposed sample quality-based updates on the parameters.

10 Conclusion and Future Work

In this paper, we show the efficacy of the framework where data is automatically labeled and both NLU and NLG models learn with quality-weighted weak supervision so as to account for the individual data quality. Most importantly, we show that not only is the two-step training process useful in improving the model, it yields decent quality text. This work serves as a starting point for weakly-supervised learning in natural language generation, especially for topics related to instance-based weighting approaches.

For future work, we hope to extend on the framework and propose ways with which it can be incorporated into existing text annotation systems.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 248 "Foundations of Perspicuous Software Systems". We sincerely thank the anonymous reviewers for their insightful comments that helped us to improve this paper.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. 2017. Learning the structure of generative models without labeled data. *Proceedings of machine learning research*, 70:273.

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained decoding for neural nlg from compositional representations in task-oriented dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 831–844.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Emanuele Bugliarello, Sabrina J Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of english than into it: Measuring neural translation difficulty by cross-mutual information. *arXiv preprint arXiv:2005.02354*.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Ernie Chang, David Adelani, Xiaoyu Shen, and Vera Demberg. 2020a. Unsupervised pidgin text generation by pivoting english data and self-training. In *In Proceedings of Workshop at ICLR*.

Ernie Chang, Jeriah Caplinger, Alex Marin, Xiaoyu Shen, and Vera Demberg. 2020b. Dart: A lightweight quality-suggestive data-to-text annotation tool. In *COLING 2020*, pages 12–17.

Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui. Su. 2021a. Neural data-to-text generation with lm-based text augmentation. *EACL 2021*.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021b. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *EACL 2021*.

Zhiyu Chen, Harini Eavani, Yinyin Liu, and William Yang Wang. 2020. Few-shot nlg with pre-trained language model. *ACL*.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642.

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2017. Fidelity-weighted learning. *arXiv preprint arXiv:1711.02799*.

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2018. Fidelity-weighted learning. In *International Conference on Learning Representations*.

Ondřej Dušek and Filip Jurcicek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2020. Trove: Ontology-driven weak supervision for medical entity classification. *arXiv preprint arXiv:2008.01972*.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. *arXiv preprint arXiv:2004.06577*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. Gpt-too: A language-model-first approach for amr-to-text generation. *arXiv preprint arXiv:2005.09123*.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017b. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Raheel Qader, François Portet, and Cyril Labbé. 2019. Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 552–562.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Esteban Safranchik, Shiying Luo, Stephen H Bach, Elaheh Raisi, Stephen H Bach, Stephen H Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, et al. 2020. Weakly supervised sequence tagging from noisy rules. In *AAAI*, pages 5570–5578.

Martin Schmitt and Hinrich Schütze. 2019. Unsupervised text generation from structured data. *arXiv preprint arXiv:1904.09447*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

Xiaoyu Shen, Ernie Chang, Hui Su, Jie Zhou, and Dietrich Klakow. 2020. Neural data-to-text generation via jointly learning the segmentation and correspondence. In *ACL 2020*. https://www.aclweb.org/anthology/2020.acl-main.641.pdf.

José GC de Souza, Michael Kozielski, Prashant Mathur, Ernie Chang, Marco Guerini, Matteo Negri, Marco Turchi, and Evgeny Matusov. 2018. Generating e-commerce product titles and predicting their quality. In *INLG*, pages 233–243.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2019. Dual supervised learning for natural language understanding and generation. *arXiv preprint arXiv:1905.06196*.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2020. Towards unsupervised language understanding and generation by joint dual learning. *ACL*.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Raphael Tang, Yao Lu, and Jimmy Lin. 2019. Natural language generation for effective knowledge distillation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 202–208.

Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1795–1807.

Hongmin Wang. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 420–425.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the*

Twenty-Fifth International Joint Conference on Artificial Intelligence, pages 2993–2999.

Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.